# Equivalent Key Frames Selection Based on Iso-Content Principles

Costas Panagiotakis, Anastasios Doulamis and Georgios Tziritas

***Abstract*— We present a key frames selection algorithm based on Iso-Content Distance, Iso-Content Error and Iso-Content Distortion principles. Under Iso-Content Distance principle, the estimated key frames are equidistant in video content. Under Iso-Content Distortion principle, the intra-clusters distance derived by the key frames are equal-sized. Under Iso-Content Error principle, the polygonal curve produced by the linear interpolation of the successive key frames will approximate the content curve over the time. In addition, two automatic approaches for defining the most appropriate number of key frames are proposed by exploiting supervised and unsupervised content criteria. Experimental results and the comparisons with existing methods from literature on large dataset of real-life video sequences illustrate the high performance of the proposed schemata.**

***Index Terms*— Key Frame selection, video-content representation, video summarization, equal distance principle.**

## I. INTRODUCTION

The traditional representation of video files as a sequence of numerous consecutive frames, each of which corresponds to a constant time interval, while being adequate for viewing a file in a movie mode, presents a number of limitations for the new emerging multimedia services such as content-based search, retrieval, navigation and video browsing. Thus a non-sequential (non-linear) video content representation has to be provided, by extracting a small but meaningful information of the visual content. Therefore, it is important to segment the video into homogenous segments in content domain and then to describe each segment by a small and sufficient number of frames (key frames) [1]. Key frames can be defined as a subset of a video sequence that can represent the video visual content as close as possible, with a limiting number of frame information [2]. Usually, the key frame extraction algorithms assume that the video file has been segmented into shots and then extract within each shot a small number of representative frames (key frames). A shot can be defined as a sequence of frames that are or appear to be continuously captured from the same camera. Ideally, a shot can encompass pans, tilts, zooms or any other camera effects [3].

Often, before applying a video summarization algorithm, appropriate visual features are extracted from each video file so as to represent its content towards a human-based semantic framework [4]. Visual content descriptors like color-texture descriptors, color-edge histograms, motion vectors have been used in key frames selection methods [5]. In order to develop an open and interoperable framework for multimedia content description, the ISO MPEG group launched the MPEG-7 standard, providing an XML (eXtended Markup Language)-based language, called Description Definition Language (DDL), for visual content representation [6].

Key frames selection approaches can be classified into cluster-based methods, energy minimization-based methods and sequential methods. The clustering techniques [1], [7] take all the frames of a shot together and classify them according to their content similarity. Then, key frames are determined as the representative frames of a cluster. The disadvantage of these approaches is that the temporal information of a video sequence is omitted. The energy minimization based methods [8] extract the key frames by solving a rate-constrained problem. These methods are generally computational expensive, since they use iterative techniques to perform minimization. The sequential methods [9] consider a new key frame when the content difference from the previous key frame exceed a predefined threshold that is determined by the user. Three approaches for video summarization have been proposed in [10]. These approaches examine the temporal variation of the feature vector trajectory or minimize a cross-correlation criterion, so that the most uncorrelated frames in feature content domain are considered as the most appropriate key frames. Recently, dynamic programming techniques have been proposed in the literature, such as the MINMAX approach of [11] to extract the key frames of a video sequence. In this work, the problem is solved optimally in $O(N^2 \cdot K_{max})$, where $K_{max}$ is related to the rate-distortion optimization. In [12], a video is represented as a complete undirected graph and the normalized cut algorithm is carried out to globally and optimally partition the graph into video clusters. The resulting clusters form a directed temporal graph and a shortest path algorithm is proposed for video summarization.

Most of the above mentioned approaches address the video summarization problem focusing either on a restricted video content, ignoring temporal variation, minimizing metric criteria on feature domain, or applying simple clustering-based techniques. On the contrary, in this paper, video summarization is performed by the use of an innovative computational geometry algorithm [13], which equally partitions the *content curve* of a video sequence resulting in key frames that are *equivalent* in the content domain under any type of video content description. We propose three general principles based on this algorithm that can be used in key frames selection:



C. Panagiotakis and G. Tziritas are with the Department of Computer Science, Univercity of Crete, Heraklion, Greece, P.O. Box 2208. A. Doulamis is with the Department of Electronic Computer Engineering, Technical University of Crete, Chania, Greece, P.O. Box 73132. E-mails:{*cpanag,adoulam,tziritas*}*@csd.uoc.gr*.

- Applying the equipartition algorithm directly on content description, we get equidistant key frames in the sense of video content, named Iso-Content Distance principle.
- Under Iso-Content Distortion principle the selected key frames minimizes a global distortion criterion providing at the same time equal distortions per key frame cluster.
- Alternatively, similar with the general polygonal approximation problem [14], the key frame selection problem can be reduced to equipartition (EP) problem. This way, the selected key frames are not selected just as a result of a criterion minimization scheme, but they are generated as points that yield equivalent partitioning of the content curve in regions with equal approximation errors. This is the Iso-Content Error principle.

The proposed method can be used under other content based criteria provided equidistant key frames. Under any of the proposed principles, the number of key frames can be computed automatically under supervised or unsupervised content based criteria. The main contribution of this work, is to address the problem of video summarization from different views, the proposed Iso-Distance, and Iso-Error principles, that take into account the equivalent property of the key frames under any type of content description. It holds that under Iso-Content Error and Iso-Content Distortion principles (see Sections II-B and II-C) the maximum frame distortion between the original sequence and its reconstruction is almost minimized, as the distortion is shared between all the segments (equal distortions). This type of criterion (maximum distortion minimization) is found to be a good metric that matches the subjective perception of the distortion [11], [15].

The rest of the paper is organized as follows: Section II gives the problem formulation describing the proposed Iso-Content Distance, Error and Distortion principles. Section III presents the proposed key frames selection algorithm and the visual content descriptors. Section IV describes two proposed approaches for the selection of key frames number. The experimental results and comparisons with the existing key frames selection methods are given in Section V. Finally, conclusions and discussion are provided in Section VI.

## II. Problem formulation

Let us assume a video shot of $N$ frames duration and that for each frame of the shot, several descriptors have been extracted and included in a vector. Let us denote as $P$ a set which includes all vectors of the sequence. Let $M$ be the number of the selected key frames and $t'_i \in [0,1], i \in \{1,\cdots,M\}$ be the ith selected key frame under the normalized time space. In the proposed method, the first and the last key frame are selected as the first and the last frame of the shot sequence, that is ($t'_1 = 0, t'_M = 1$). Therefore, the goal of the proposed method is to compute the remaining $M-2$ key frames $t'_i, i \in \{2,\cdots,M-1\}$, under the constraint that are equidistant in the sense of the used semimetric function $g(x,y)$ [14]. That is,

$$r = g(t'_{i-1}, t'_i) = g(t'_i, t'_{i+1}),\ i \in \{2,\cdots,M-1\}, \qquad (1)$$

where $r$ denotes the length of each equal chord. Therefore, the set of key frames are selected under any predefined content description having equivalent property on video content descriptors.

The assumption of using the first and last frame of a shot as two starting key frames may not be appropriate when the shot segmentation failed. This could be caused due to factors such as fading. However, there are methods that can automatically recognize such effects improving the shot segmentation [16].

### A. Iso-Content Distance Principle

Under Iso-Content Distance principle, the content distances between two successive key frames should be equal. Thus, under the definition of Section II, we have to compute $M-2$ sequential key frames $t'_i$ under the constraint: $r = d(t'_{i-1}, t'_i) = d(t'_i, t'_{i+1}),\ i \in \{2,\cdots,M-1\}$, where $d(x,y),\ x,y \in [0,1]$ denotes the semimetric distance function. Several distances $d(x,y)$ can be used, like Euclidean, Manhattan, $\chi^2$, depending on content descriptors' space. If there are several solutions, the one with the maximum chord length $r$ is selected, since this solution is the best approximation of the content curve.

### B. Iso-Content Error Principle

Under Iso-Content Error principle, the content error distances between two successive line segments of the polygonal curve $P'$ produced by the linear interpolation of the successive key frames will be equal.

Different error criteria have been proposed for polygonal approximation problems. One of the most used is the tolerance zone criterion [17]. Under this criterion, the error between the line segment and the corresponding (polygonal) subcurve of $P$, $S$ is defined as the maximum distance between the line segment and each point on the subcurve $S$. Another frequently used error criterion is the local integral square error (LISE) [18]. Under this criterion, the error between the line segment and $S$ is defined as the sum of squared Euclidean distances from each vertex point of subcurve $S$. The error under tolerance zone is independent from the number of frames between the successive key frames, while the LISE criterion depends on this number. Thus, under LISE criterion the key frames are usually more equally spaced in time than under tolerance zone criterion. Tolerance zone criterion is suitable for applications where the duration of the time intervals between the key frames can be entirely ignored. If there are several solutions, the one with the minimum error is selected.

### C. Iso-Content Distortion Principle

Under Iso-Content Distortion principle, the distortions between two pairs of key frames, $\bar{d}(t'_{i-1}, t'_i) = \bar{d}(t'_i, t'_{i+1}),\ \forall i \in \{2,\cdots,M-1\}$, should be equal. We consider the following definition for distortion: The distortion $\bar{d}(t'_i, t'_{i+1})$ is defined as the sum of minimum content distances of the frames $t_j$, $t'_i \le t_j \le t'_{i+1}$ and the two successive key frames $t'_i,\ t'_{i+1}$,

$$\bar{d}(t'_i, t'_{i+1}) = \sum_{j=t'_i}^{t'_{i+1}} \min(d(t'_i, t_j), d(t_j, t'_{i+1})). \qquad (2)$$

This definition is similar with that of distortion used by Lee and Kim [8] and which is given by the following equation,

$$\bar{d}(K,B) = \sum_{i=1}^{M} \int_{b_i}^{b_{i+1}} d(t,t_i')dt. \quad (3)$$

In Equation (2), we have not used the breakpoints ($B = \{b_0, \cdots, b_M\}$) as in Equation (3), since their meaning is included in the successive key frames. If we define the total distortion as the maximum of the corresponding distortions $\max_{i\in\{1,2,\cdots,M-1\}} \bar{d}(t_i', t_{i+1}')$ (as is defined by the term maximum frame distortion between the original sequence in [11]), similar to the total polygonal error of Section II-B, then almost optimal solutions are achieved using the proposed schema. If there are several solutions, the one with the minimum distortion is selected.

## III. Key Frames Selection Algorithm

The straightforward implementation of the EP method provides directly $M$ key frames. The number of key frames $M$ can be given by the user or can be estimated automatically by terminating the EP algorithm when the estimated "distortion" exceeds a predefined error, similar with the problem of minimum number of segments ($min - \#$)[1] [14]. Both cases are solved in $O(M \cdot N^2)$ steps thanks to the property of the method that it solves the problem for values less than $M$ without additional cost [14]. In the aforementioned cost, we have not included the requirements in function $g(t_k, t_l)$ computation, which is usually about $O(n \cdot N^2)$. An important algorithm property is that the computation cost is independent of content curve dimension $n$, that coincides with the feature vector space dimension. This means that it is independent of content descriptors selection.

The input of the proposed method is the number of key frames $M$. In addition, it needs the values of symmetric matrix $g(t_k, t_l), k, l \in \{1, 2, \cdots, N\}$ of distortions. The detailed description of the algorithm used can be found in [14]. It is an iterative method. Thus, when it is executed for $M$ segments, it uses the precomputed results for $M-1$ segments. In each iteration step $l$, the algorithm computes the zero level curves $L_l$ by the $L_{l-1}$. According to our analysis [13], the equipartition problem (EP) always admits at least one solution.

The proposed method can be executed under any choice or combination of audio/visual content descriptors. However, the selected key frames are related with the used content description, so we have to choose appropriate descriptors. On this framework, we use MPEG-7 visual descriptors [5], like the Color Layout Descriptor (CLD), a low cost and compact descriptor, which suffices to describe smoothly the changes in visual content (mainly color and motion variations) of a shot. We used the following semimetric function $D$ to measure the content distance of two CLDs, $\{DY, DCb, DCr\}$ and $\{DY', DCb', DCr'\}$, $D = \sqrt{\sum_i (DY_i - DY_i')^2} + \sqrt{\sum_i (DCb_i - DCb_i')^2} + \sqrt{\sum_i (DCr_i - DCr_i')^2}$, where $(DY, DCb, DCr)$ represent the i$^{th}$ DCT coefficients of the respective color components.

[1] $min - \#$ is used for the problem of finding the minimum number of segments that gives error lower than the given error (at polygonal approximation).

## IV. Selection of the Key Frame Number

The selection of key frame number ($M$) could be done by the user to fit his/her specific preferences and information needs. However, it is crucial to develop an mechanism able to automatically estimate the most appropriate number of key frames $M$ used in the summarization process. Two different approaches are proposed in this paper for the automatic calculation of $M$. The first scheme exploits a *supervised learning process*, while the other is based on an *unsupervised schema*.

### A. Supervised approach

The supervised approach can be used under the three proposed principles (see Section II). In particular, we initially compute the maximum content distance over a learning set of shots each of which is described using the adopted CLD descriptor of MPEG-7. This distance refers to the pair of shot frames whose distance is the maximum distance over all pairs of frames in the shot. The CLD descriptors are scale invariant, so we can compute the mean maximum shot content distance $D_e$ by averaging over the computed maximum content distances. $D_e$ was estimated $589.98$ getting the mean value of 200 shots.

The steps of the algorithm under Iso-Content Distance principle are described hereafter. The user gives a percentage of $D_e$, $a \cdot D_e$ (e.g. $a = 30\%$) instead of number $M$. If the maximum shot content distance is lower than the given $a \cdot D_e$, then $M$ is set to 2, i.e., only the first and the last frame of the shot are selected as key frames. Otherwise, EP algorithm is iteratively executed computing in each iteration $l$, $l+1$ key frames and the estimated equal content distance $r_l$. The algorithm terminates when the $r_l$ is less than $a \cdot D_e$. At the same time, the key frames are provided. Under Iso-Content Error or Iso-Content Distortion principles, the procedures are exactly the same as under Iso-Content Distance principle. However, in these cases the given $a \cdot D_e$ should be normalized in order to be comparable with the content approximation errors $r_l$. Thus, under tolerance zone criterion, we have just to multiply it by a constant $c$ using $c \cdot a \cdot D_e$, while under LISE criterion we have to square $(c \cdot a \cdot D_e)^2$, using a higher $c$. By our experiments, $c = 0.4$ and $c = 2$ yield good results under tolerance zone criterion and LISE, respectively.

Parameter $a$ is closer to the human perception, as far as the visual content is concerned in contrast to the number $M$, since it expresses the percentage of the expected maximum content variation. If the shot is rich in content, $M$ will be high, otherwise $M$ will be low, under a constant percentage of $D_e$.

### B. Unsupervised approach

According to the unsupervised approach, $M$ is provided without any user interaction estimating if the selected key frames at iteration $l$ of the algorithm suffice to approximate the content curve.

- Under Iso-Content Distance principle, the function $Q_l$ (see Equation (4)) is a measurement of the distance

between the content curve and the key frames approximation of the content curve using the $l+1$ key frames on level $l$ of EP algorithm.

$$Q_l = \sum_{i=1}^{N-1} g(t_i, t_{i+1}) - l \cdot r_l, \quad l \in \{2, 3, \cdots, N-1\} \quad (4)$$

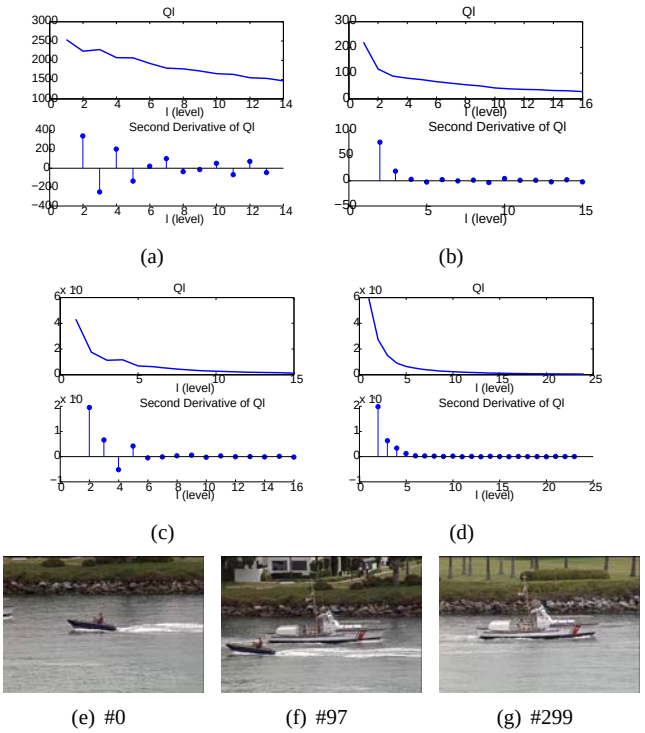
In the case of $l = 1$, $Q_1$ is defined by $\sum_{i=1}^{N-1} g(t_i, t_{i+1}) - g(t_1, t_N)$. $\sum_{i=1}^{N-1} g(t_i, t_{i+1})$ and $l \cdot r_l$ are the $P$ and $P'$ curve lengths, respectively. Thus, the better content description of $P$ is achieved by the set $P'$, when the two lengths are closer to each other.

- Under Iso-Content Error principle and tolerance zone criterion, as $Q_l$ we use directly the error of the polygonal approximation, $Q_l = r_l$. Under Iso-Content Error principle and LISE criterion, as $Q_l$ we use the error of the polygonal approximation multiplied by the number of segments, $Q_l = l \cdot r_l$. This is done since the error under LISE criterion is locally summed and the number of coefficients are inversely proportional to the number of segments or the level of EP algorithm.
- Under Iso-Content Distortion principle, as $Q_l$ we use the distortion multiplied by the number of segments, $Q_l = l \cdot r_l$ similar with previous case.

$Q_l$ is usually decreasing as $l$ increasing, $Q_l \geq 0$. $Q_l$ has characteristics of a convex function, that is, if we smooth it we will get a convex function. Therefore, we have to introduce a new criterion instead of minimum of this function. Thus, we propose to select the appropriate level $l$ so that the numerical approximation of the second derivative of $Q_l$, $\ddot{Q}_l$, is maximized.

$$\ddot{Q}_l = Q_{l+1} + Q_{l-1} - 2 \cdot Q_l, \quad l \in \{2, 3, \cdots, N-2\} \quad (5)$$

This is due to the fact that the second derivative expresses a measure of the curvature of the content curve.






(e) #0 (f) #97 (g) #299

Fig. 1. The functions $Q_l$, $\ddot{Q}_l$ under **(a)** Iso-Content Distance, **(b)** Iso-Content Error and tolerance zone, **(c)** Iso-Content Error and LISE **(d)** Iso-Content Distortion in coast sequence. The proposed number of key frames were the same under any proposed principle, $M = 3$. **(e), (f), (g)** The proposed key frames were exactly the same under using Iso-Content Distance or Iso-Content Error principle and tolerance zone criterion, $K = \{0, 97, 299\}$.

## V. Experimental Results

### A. Dataset content description

In this section, the experimental results of the proposed algorithm and comparisons to other algorithms are presented. We have tested the proposed algorithm on a data set consisting of more than 250 real life video sequences. The most of them are videos from athletic meetings, like pole vault, high jump, triple jump, long jump, running and hurdling. Moreover, we have used the widely known as MPEG test sequences like coast sequence, the table tennis sequence, hall monitor sequence, etc. The number of frames per shot were mainly varied between 200 and 500. Figs. 6, 7, 8 and 9 illustrate the sequences that we used in the article. A typical processing time for the execution of the proposed EP algorithm, when the shot contains 300 images (e.g. coast MPEG sequence) and $M = 10$, is 4 to 5 seconds depending on the used principle.

### B. Evaluation of the Proposed Scheme

The results of the proposed schemata under several video sequences are presented hereafter. More specifically, in coast sequence (see Fig. 6) three key frames are extracted to represent the shot visual content. This number is obtained by the application of the algorithm described in Section IV-B and coincides with the number of phases involved in this scene. More specifically, the sequence shows two boats that are moved across a river. First the camera tracks a single black boat. Next, a white boat is appearing. Finally, the black boat disappears from the scene and the camera tracks the white boat. Fig. 1 illustrates results of the four proposed schemata in the coast sequence. The selected key frames represent the coast sequence very well, showing three characteristics phases, the single black boat, the white black boat and both of them. Concerning an objective criterion, the proposed method is optimal in the sense of equal partition of video content curve under the used metric (e.g. polygonal approximation error, distance, Distortion).

Fig. 2 illustrates the results of the four proposed schemata in pole vault sequence, using five key frames, representing the shots well. To derive a fair comparison among all the four proposed schemata, the same number of key frames are used in Fig. 2, which coincides with the minimum value of the optimal number obtained by the application of the algorithm of Section IV-B for the four proposed schemata. On pole vault sequence, the camera tracks the athlete. The visual content is mainly varied at the end of the sequence, when the athlete is in jumping phase. We have observed that, under Iso-Content Error principle the key frames are usually selected close to "corners", special points of content curve, like where

the curvature is changed, while under Iso-Content Distortion principle the better representative frames of their cluster are selected. Moreover, the selected key frames under Iso-Content Distance principle or Iso-Content Error principle and tolerance zone criterion don't take the duration between the selected key frames into account, while the other two approaches combine the duration with the content variation. Thus, it holds that, when the content suddenly changes, more key frames are selected under the two aforementioned schemata than under the other two schemata. The four proposed schemata describe well the visual content of a foreman subshot (Fig. 4) and 140 shot of docon.mpg [12] (see Fig. 5) using five and three key frames, respectively.

Figs. 2(a), 2(g), 2(m) and 2(s) illustrate the surfaces $g(x, y)$ under the proposed principles. The deep blue colors correspond to close to zero values. The deep red colors correspond to the highest values of $g(x, y)$. The estimated solution is projected on $g(x, y)$ with cycles. The $L_l$ curves are projected on $g(x, y)$, with gray colors, at both sides of diagonal $x = y$. If more than one solutions are appeared, the selected solution points are drawn with large cycles. It can be observed that under Iso-Content Distortion or Iso-Content Error principle and LISE criterion the derived distance matrix $g(x, y)$ and the $L_l$ curves are smoother than the corresponding distance matrix and curves under the other two principles.

### C. Comparison to Other Algorithms

In the following, we compare the results obtained by the application of the proposed algorithm with other video summarization techniques presented in the literature. More specifically, we have compared the performance of the proposed method with the method of [10] that estimates the most appropriate key frames in a video sequence by minimizing a cross correlation criterion and the work of [19] which handles the summarization problem as an interpolation problem. In the first approach the most uncorrelated frames in content domain, as expressed by a feature description, are selected as key frames, since they represent all possible variations of content properties. The other approach initially plots the norm of the feature vector trajectory through time and then estimates as key frames those points that minimize the trajectory interpolation error as being linearly approximated by the selected points. Both techniques are computational demanded such as the NP-complete problems and thus optimized algorithms are required to get solutions close to the optimal ones. A genetic algorithm has been adopted to quickly estimate the optimal frame indices in both cases. Moreover we have compared the proposed work with the MINMAX method [11] and a graph modeling approach [12]. In [11], the reconstructed video sequence is obtained from the set of key frames, by substituting the missing frames with the most recent frame that belongs to set $K$. According to MINMAX criterion, the set of key frames is selected in order to minimize the maximum frame distortion between the original sequence and its reconstruction (using the set of key frames). In [12], scene change detection and video summarization is done by modeling the evolution of a video through a temporal graph. They employ a global criterion, normalized cut, to optimally decompose the graph into subgraphs (clusters). Ideally shots in a cluster will share similar video content after the partitioning. Finally, based on the sorted order of shots, they discard one subshot at a time until the desired skim ratio is reached in order to maintain the content balance.

The key frames extracted for the pole vault sequence using the interpolation method [19] and the genetic implementation of [10] are depicted in Fig. 3. In this case and in order to obtain a fair comparison, the same number of frames as the one derived from the adopted automatic process is used ($M = 5$). It is still evident that the proposed approach in all principles outperforms the compared methods. In particular, genetic searching technique used in [10] cannot capture the temporal periodicity of the pole vault sequence (the scene returns to the pillows after the athlete jump). Instead, the work [19] exploits such a temporal variation, but with a simple linear interpolation of the visual content.

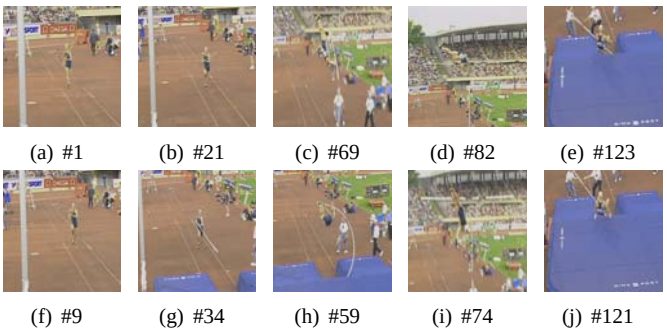
The key frames extracted for the foreman subshot sequence (150-249 frames) using MINMAX method with $K_{max} = 50$ and the four proposed schemata are depicted in Fig. 4. The MINMAX method does not give any key frame during the camera rotation, while its last two key frames have almost equal content.

The key frames extracted for the 140 shot of docon.mpg using a graph modeling approach [12] and the four proposed schemata are depicted in Fig. 5. It holds that the key frames of the proposed schemata describe well the visual content showing the three characteristics phases of the sequence: the first hero, the second hero and the background, while the graph modeling approach misses the last one.

It should be mentioned that the adopted works that are used for comparisons are not the simpler summarization techniques but use some optimal criteria. We make such selection in order to demonstrate the efficiency of the proposed scheme under arduous situations. The physical reason that leads our method to outperform the compared ones (in the experimental results) is due to the equal-distance selection process in content domain that is adopted in our work.

## VI. Conclusions

Four key frames selection schemata are described based on equipartition problem. The first schema, called Iso-Content Distance, extracts as key frames the ones that are equidistant in video content. Under the Iso-Content Error principle, the content curve derived by the linear interpolation of the successive key frames will approximate the content curve over the time based on LISE (second schema) or tolerance zone criterion (third schema). Under the fourth schema, called Iso-Content Distortion, the frames clusters size derived by the key frames is equalized (size). Thus, the selected key frames have different properties according to the used principle. However, in any case, the key frames are equivalent on content video summarization. Each key frame has the same significance under the used principle. At the same time the maximum frame distortion between the original sequence and its reconstruction is almost minimized, which is a good metric that matches the



(a) #1 (b) #21 (c) #69 (d) #82 (e) #123

(f) #9 (g) #34 (h) #59 (i) #74 (j) #121

Fig. 3. Results of interpolation [19] and genetic method [10] in pole vault shot using five key frames. **(a),** $\cdots$**, (e)** The selected key frames under interpolation method. **(f),** $\cdots$**, (j)** The selected key frames under genetic method.

subjective perception of the distortion. Concerning, the number of key frames, two automatic approaches are proposed each exploiting supervised or unsupervised content based criteria. The proposed method can be applied by any type of descriptors extracting the key frames without any changes on the proposed algorithm. In the framework of this work, we have used the Color Layout Descriptor (CLD) of MPEG-7 standard to guarantee interoperability.

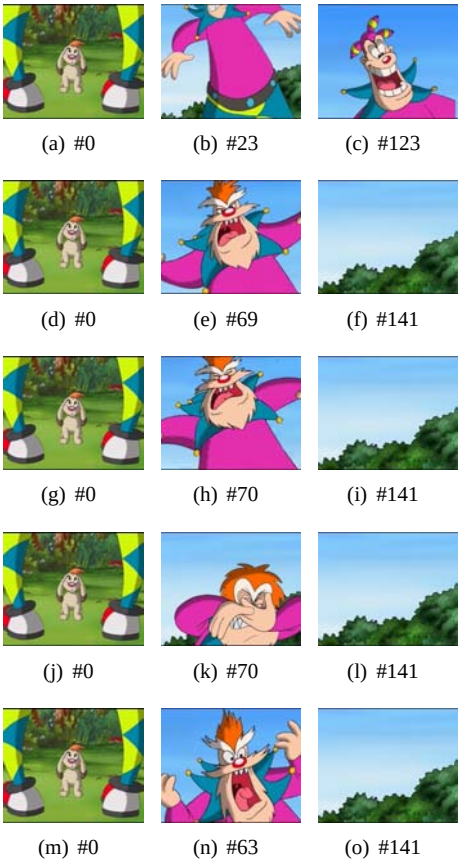
Experimental results on a large data set of real-life video sequences have been conducted to demonstrate the efficiency of the proposed schemata as far as the visual content representation is concerned. In all cases, the appropriate number of key frames as obtained by the proposed automatic processes is close to the human's perception with respect to the content fluctuation. Comparisons with other optimal video summarization techniques, such as the works of [10]–[12], [19], indicate the high performance of the adopted method in relation to the compared ones. A possible extension of the proposed methodology may include the addition of more audio/visual descriptors or the using of other principles. Moreover, a weighted combination of the proposed principles can be examined providing key frames that combine the principles' properties. We can extend the proposed method in order to solve the problem of video summarization (selection of key frames in a sequence of shots). First, we can compute the content change per shot, using the content distance function $g(x, y)$. Then we can estimate the number of key frames per shot, so that the ratio between number of key frames of shots, will be almost equal with the ratio of their content changes.

## ACKNOWLEDGMENT

This research was partially supported by the Greek PENED 2003 project. The authors would like to thank the researchers of LIS (Image and Signal processing Lab) at Grenoble, Emmanuel Ramasso, Michèle Rombaut and Denis Pellerin for the data exchange.

## REFERENCES

[1] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based onunsupervised cluster-validity analysis," *IEEE Trans. On Circuits And Systems For Video Tech.*, vol. 9, no. 8, pp. 1280–1289, 1999.



(a) #0 (b) #23 (c) #123

(d) #0 (e) #69 (f) #141

(g) #0 (h) #70 (i) #141

(j) #0 (k) #70 (l) #141

(m) #0 (n) #63 (o) #141

Fig. 5. **(a), (b), (c)** The selected key frames under Graph Modeling approach [12] in a shot of docon.mpg [12]. **(d), (e), (f)** The selected key frames under Iso-Content Distance principle. **(g), (h), (i)** The selected key frames under Iso-Content Error and LISE criterion. **(g), (k), (l)** The selected key frames under Iso-Content Error and tolerance zone criterion. **(m), (n), (o)** The selected key frames under Iso-Content Distortion principle.

[2] M. Yeung and B.-L. Yeo, "Video visualization for compact presentation and fast browsing ofpictorial content," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 7, no. 5, pp. 771 – 785, 1997.

[3] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods." *IEEE Trans. Circuits Syst. Video Techn.*, vol. 10, no. 1, pp. 1–13, 2000.

[4] Y.-P. Tan, S. R. Kulkarni, and P. J. Ramadge, "A Framework For Measuring Video Similarity And Its Application To Video Query By Example," 1999.

[5] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. On Circuits And Systems For Video Tech.*, vol. 11, no. 6, pp. 703–715, 2001.

[6] MPEG-7-Group, "Mpeg-7 applications document," *ISO/IEC/JTCI/SC29/WGI //N2462*, 1998.

[7] A. Girgensohn and J. S. Boreczky, "Time-constrained keyframe selection technique," *Multimedia Tools and Applications*, vol. 11, no. 3, pp. 347–358, 2000.

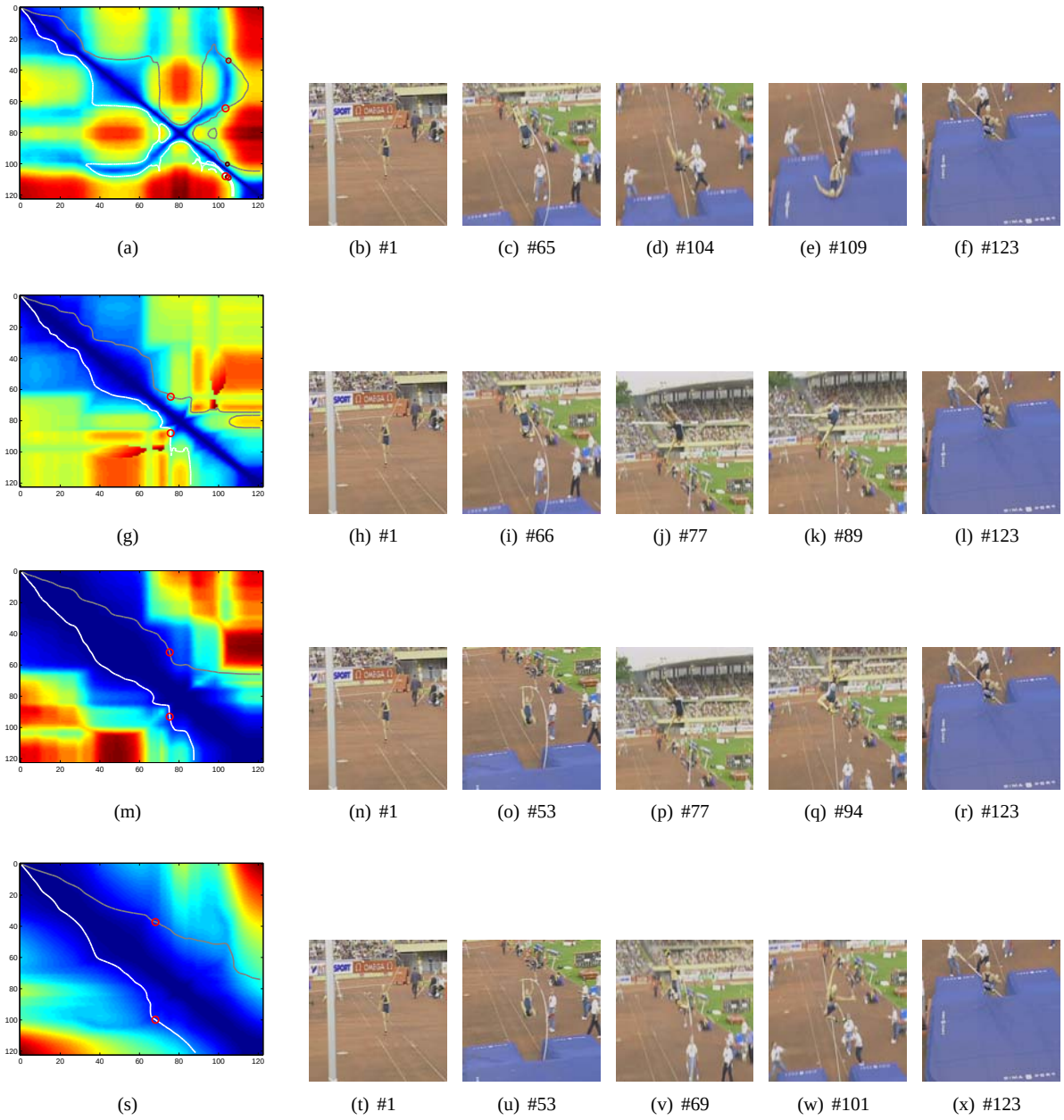
[8] H.-C. Lee and S.-D. Kim, "Iterative key frame selection in the rate-constraint environment," *Signal Processing: Image Communication*, vol. 18, pp. 1–15, 2003.

[9] J. Vermaak, P. Perez, and M. Gangnet, "Rapid summarization and browsing of video sequences," in *British Machine Vision Conf.*, 2002.

[10] Y. Avrithis, A. Doulamis, N. Doulamis, and S. Kollias, "A stochastic framework for optimal key frame extraction from mpeg video databases," *Journal of Computer Vision and Image Understanding,*, vol. 75, no. 4, pp. 3–24, 1999.

[11] Z. Li, G. Schuster, and A. Katsaggelos, "Minmax optimal video summarization," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 15, no. 10, pp. 1245 – 1256, 2005.

[12] ——, "Video summarization and scene detection by graph modeling,"



(a) (b) #1 (c) #65 (d) #104 (e) #109 (f) #123

(g) (h) #1 (i) #66 (j) #77 (k) #89 (l) #123

(m) (n) #1 (o) #53 (p) #77 (q) #94 (r) #123

(s) (t) #1 (u) #53 (v) #69 (w) #101 (x) #123

Fig. 2. Results of the four proposed schemas in pole vault shot using five key frames. The estimated solution and the $l_t$ curves are projected on $g(x, y)$ under **(a)** Iso-Content Distance, **(g)** Iso-Content Error and LISE criterion, **(m)** Iso-Content Error and tolerance zone criterion and **(s)** Iso-Content Distortion principle. **(b),** $\cdots$**, (f)** The selected key frames under Iso-Content Distance principle. **(h),** $\cdots$**, (l)** The selected key frames under Iso-Content Error and LISE criterion. **(n),** $\cdots$**, (r)** The selected key frames under Iso-Content Error and tolerance zone criterion. **(t),** $\cdots$**, (x)** The selected key frames under Iso-Content Distortion principle.

*IEEE Trans. Circuits Syst. Video Techn.*, vol. 15, no. 2, pp. 296 – 305, 2005.

[13] C. Panagiotakis, G. Georgakopoulos, and G. Tziritas, "On the curve equipartition problem: a brief exposition of basic issues," in *European Workshop on Computational Geometry*, 2006.

[14] C. Panagiotakis and G. Tziritas, "Any dimension polygonal approximation based on equal errors principle," *Pattern Recogn. Lett.*, vol. 28, no. 5, pp. 582–591, 2007.

[15] H. Sundaram and S.-F. Chang, "Constrained utility maximization for generating visual skims," in *IEEE Workshop Content-Based Access of Image and Video Library*, 2001, pp. 124 – 131.

[16] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 16, no. 1, pp. 82–91, 2006.

[17] H. Imai and M. Iri, "Polygonal approximations of a curve (formulations and algorithms)," *Computational Morphology*, pp. 71–86, 1988.

[18] K.-L. Chung, W.-M. Yan, and W.-Y. Chen, "Efficient algorithms for 3-d polygonal approximation based on lise criterion," *Pattern Recognition*, vol. 35, pp. 2539–2548, 2002.

[19] N. Doulamis, A. Doulamis, and K. Ntalianis, "An optimal interpolation-based scheme for video summarization." IEEE Inter. Conf. on Multimedia and Expo (ICME), 2002.

(a) #1 (b) #29 (c) #80 (d) #94 (e) #99

(f) #1 (g) #32 (h) #45 (i) #64 (j) #100

(k) #1 (l) #29 (m) #45 (n) #66 (o) #100

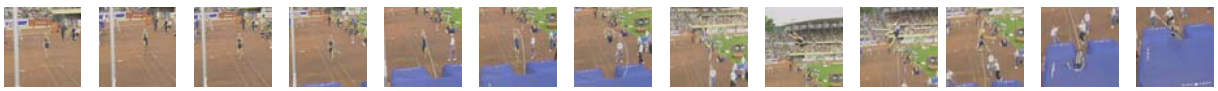
(p) #1 (q) #32 (r) #40 (s) #62 (t) #100

(u) #1 (v) #29 (w) #49 (x) #69 (y) #100

Fig. 4. **(a),** $\cdots$**, (e)** The selected key frames under MINMAX method [11] in second subshot of foreman sequence. **(f),** $\cdots$**, (j)** The selected key frames under Iso-Content Distance principle. **(k),** $\cdots$**, (o)** The selected key frames under Iso-Content Error and LISE criterion. **(p),** $\cdots$**, (t)** The selected key frames under Iso-Content Error and tolerance zone criterion. **(u),** $\cdots$**, (y)** The selected key frames under Iso-Content Distortion principle.

(a) #0 (b) #30 (c) #60 (d) #90 (e) #120 (f) #150 (g) #180 (h) #210 (i) #240 (j) #270 (k) #299

Fig. 6. The coast sequence which contains 300 frames.

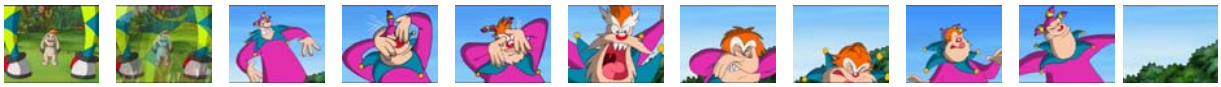


(a) #1 (b) #10 (c) #20 (d) #30 (e) #40 (f) #50 (g) #60 (h) #70 (i) #80 (j) #90 (k) #100 (l) #110 (m) #123

Fig. 7. The pole vault sequence which contains 123 frames.

(a) #1 (b) #10 (c) #20 (d) #30 (e) #40 (f) #50 (g) #60 (h) #70 (i) #80 (j) #090 (k) #100

Fig. 8. A subshot (150-249 frames) of foreman sequence.



(a) #0 (b) #15 (c) #30 (d) #45 (e) #60 (f) #75 (g) #90 (h) #105 (i) #120 (j) #130 (k) #141

Fig. 9. The 140 shot of docon.mpg [12].